Automatic detection of gender on the blogs

Faiza Belbachir
LSSD Laboratory
USTO-MB
belbachirfaiza@yahoo.fr

Khadidja Henni
LSSD Laboratory
USTO-MB
khadidja_henni08@yahoo.fr

Lynda Zaoui
LSSD Laboratory
USTO-MB
zaoui_lynda@yahoo.fr

***Abstract*—In this paper, we are interested in defining the gender of blogger while using only texts written from bloggers. For that purpose, we offer a number of features based on specific words, which were categorized into classes. For each blog, a score is calculated based on these characteristics, thereby determining the gender of its author. The evaluation was made on a corpus of 681,288 Blogs (140 million words) tagged as men or women. In our work, this collection will be taken as a reference. The obtained results show gender detection over 82% compared to the referenced collection.**

***Keywords*-Information retrieval, blogs, gender detection, social network**

I. Introduction

Social network such as blogs, Facebook and Twitter are recently growing fast and allow users to post opinions, share feelings and ideas in different areas. Networks are a source of information to monitor user activity to find similarities between individuals, to make recommendations to improve the process of information retrieval, to extract the profile of the Internaute in order to return the most relevant documents upon request.

Much work [8], [9], [10], [11], [12], [13] are interested in users identification (gender, age, status etc ...) from the texts they write or comment their post and use this identification in several areas such as information retrieval personalized, targeted marketing, security and detection of criminals and spammers. The identification of Internaute use several techniques such as natural language processing, learning, statistics, etc.

Determining the gender of a user from texts has been studied extensively. The identification and interpretation of possible differences in linguistic styles between men and women has a significant impact in several areas such as commercial (for search engines, it is interesting to categorize users as to best meet their research survey (eg, whether women or men are for or against abortion), politics (if that politician is preferred by women or men).

In this work, we focus on blogosphere as a valuable source of this study, the fact that anyone can write a blog in any topic without any constraint of style. More blogs are available and their number is in perpetual growth (millions of blogs). However blogs are generating new challenges due to the used language by bloggers that do not follow the grammar of the language, and a lot of words are not belonging to the used dictionary.

In this paper, we first explain some related works that are interested in the detection of gender and the used feautures. We then introduce our approach based on the calculation of a score for a blog to determine the gender of its author. This score is calculated according to a set of characteristics.

These features are based on specific words to each gender and are categorized into classes. We evaluate our approach on a corpus of 681.288 blogs (140 million words) National Corpus (BNC) " [1]. We show, while evaluating on the same corpus few approaches mentioned in the state of the art, that our approach is more efficient.

II. Related works

The authors Shlomo Argamon and al. [1], were interested in determining gender in formal texts (articles, texts). Their work explores the differences between male and female writing in corpus covering a range of genders. The objective of this approach is to conclude that there are significant differences between documents male and female authors only based on the text and the used language operators (pronouns, determiners, ect ...).

For that purpose, the authors have studied a corpus of 604 documents" British National Corpus (BNC) and deduce the characteristics of each gender. They concluded that women are more susceptible to narration and dialogue and they use pronouns (personal and possessive), verbs (especially past tense), injections and quantifiers. However, men are more likely to specify or clarify. They provide more specification and they use determiners, adjectives (comparative and superlative), proper nouns, adverbs (comparative and superlative), coordinating conjunctions and numbers.

[1] Pr moshe koppel, july 2006. http:// u.cs.biu.ac.il/koppel /

The authors Nowson Scott et al. [2] used the BNC corpus and showed that there are two types of texts: formal and contextual. They introduced a measure of (contextuality/formality) text called (F_mesure) which is based on a subtraction between the frequencies of nouns, adjectives, prepositions and articles and those pronouns, verbs, adverbs, interjections. They have shown that men prefer formality in their writings (F_measure low), while women prefer contextuality (F_measure high).

The authors Shlomo Argamon and al. [3] have took into account several classes of features to predict the kind of blogger, those concerning the grammatical style (Adjective, adverb, verb, noun, etc. ..), those relating to function words, and those words are related to specific blogs such as (smily, lol ...). These researchers concluded that the use of words associated with these classes differ from one genre to another. The words of classes games, religion, politics, business, internet, article, preposition, are more often associated with men, while using words class conversation, athome, fun, romance, swearing, pronouns personal, conjunction, auxiliary verb is more specific to women. In 2007 the authors have improved their approach by expanding their classes dictionary by adding more words that characterize men and women [4], they took into account the language used by the Internet (slang, grammar, ect ...).

Sumit Goswami et al [5] chose two factors for differentiating between (male and female) bloggers: the length of sentences and use words that do not belong to the language. The analysis of blogs based on the average length of the sentence is a challenge because blogs lack editorial controls and grammar. The use of non-dictionary is not as easy as writing blogs have informal, without limitation editorial.

Arjun Mukherjee and al. [6] took into account a set of characteristics which are: F_measure [6], parsing(POS) [1], the use of a dictionary of words of content [2] the use of features such preferences [7] and the use of POS sequences [5]. The authors use SVM classifiers, and Naives Bayes classifiers.

III. Proposed approach

In our approach we were inspired by the work of Shlomo et al [4]. However, additionally to their work, we developed several important items as follow:

- Added new classes to define the blogger gender.
- Subdivided some classes into new classes to be able to precisely determine the best gender.

[2]www.hackerfactor.com/GenderGuesser.php

Classes	Words	Gender
Conversation	know, people, think, person, tell, feel, friends, talk, new, talking, mean, ask, understand, feelings, care, thinking, friend, relationship, realize, question, answer, saying.	Female
Family	years, family, mother, children, father, kids, parents, old, year, child, son, married, sister, dad, brother, moved, age, young, months, three, wife, living, college, four, high, five, died, six, baby, boy, spend, Christmas.	Female
Period	weeks, hours, days, hour, month, june, year, past, future, years, December, November, October, September, august, july, may, april, march, February, January.	Female
Work	work, working, job, trying, right, met, figure, meet, start, better, starting, try, worked, idea.	Female
PastAction_Position	saw, felt, left, called, tried, sat.	Female
Games	game, games, team, win, play, played, playing, won, season, beat, final, two, hit, first, video, second, run, star, third, shot, table, round, ten, chance, club, big, straight.	Male
Internet	site, email, page, please, website, web, post, link, check, blog, mail, information, free, send, comments, comment, using, internet, online, name, service, list, computer, add, thanks, update, message.	Male
Fun	fun, im, cool, mom, summer, awesome, lol, stuff, pretty.	Female
Funny	ill, mad, funny, weird.	Female
Conversation	know, people, think, person, tell, feel, friends, talk, new, talking, mean, ask, understand, feelings, care, thinking, friend, relationship, realize, question, answer, saying.	Female
Food/Col	food, eating, weight, lunch, water, hair, life, white, wearing, color, ice, red, fat, body, black, clothes, hot, drink, wear, blue, minutes, shirt, green, coffee, total, store, shopping.	Female
Poetic	eyes, heart, soul, pain, light, deep, smile, dreams, dark, hold, hands, head, hand, alone, sun, dream, mind, cold, fall, air, voice, touch, blood, feet, words, hear, rain, mouth.	Female
BooKs/Movies	thank, lives, earth, world.	Female
Subjective	know, people, think, person, tell, feel, friends, talk, new, talking, mean, ask, understand, feelings, care, thinking, friend, relationship, realize, question, answer, saying.	Female
Religions	god, jesus, lord, church, word, power, human, believe, given, truth, death, evil, own, peace, speak, bring, truly.	Male
Romance	forget, forever, remember, gone, true, face, spent, times.	Female
Sentiment	love, cry, hurt, wish, loved	Male
Shame	guy, shit, ass, sex, fuck, sucks.	Male
Politics	bush, president, Iraq, kerry, war, american, political, states, america, country, government, john, national, news, state, support, issues, article, michael, bill, report, public, issue, history, party, york, law, major, act, fight, poor.	Male
Music	kmusic, songs, song, band, cd, rock, listening, listen, show, favorite, radio, sound, heard, shows, sounds, amazing, dance.	Male

Table I
CLASSES OF SPECIFIC WORDS

- Defined a score, which will be calculated for a document to determine the gender of the blogger.

In the studied corpus, we considered the most frequent words. This corpus is tagged according to the blogger gender, where we classified the word into two groups: men and women. In this section we first discuss the different classes that we are going to use then we explain the developed score.

A. The classes of relevant words contents that determine the blogger gender

We improve the existent classes, subdivided some classes and added new classes such as: Period, PastAction-Position, Subjective, Funny, Shame. The contents of each used class are given as follow (show table **??**), every class is tagged or ' listed as man ' or ' listed as woman ' where the majority of the terms define the gender.

We also introduced the use of classes relative to the auxiliaries and propositions (show table II).

Classes	Words	Gender
Auxiliaries	are, been, can, could, dare, did, had, has, have, keep, may, might, must, need, ought, shall, should, used, was, were, would.	Female
Proposition	about, above, across, after, against, along, among, around, at, before, behind, below,beneath, beside, between, by, down, during, except, for, from,in, front, inside, instead, into, like, near, of, off, on, onto, top, out, outside, over, past, since, through, to, toward, under, undermeath, until, up, upon, whith, without.	Male

Table II
AUXILIARIES AND PROPOSITIONS WORDS

B. The Score that determines the gender

After defining the classes of specific words and to determining if a man or a woman writes a document, we calculate the score. This score depends on frequencies of the terms in this document and their membership in the tagged classes (Man, Woman) (see equation 1). If a document has more terms, which belong to man tagged classes than the women classes; we consider this document as a man document.

$$Cal_G(D) = \sum_{t \epsilon D \wedge t \epsilon C_G} fr(t) \quad (1)$$

With $G \epsilon \{M, F\}$ and $fr(t)$ is the frequency of term t in document (D). C_G is a tagged class G equal to (M or F). For every document, we shall have two values $Cal_M(D)$ (corresponding of terms that are into classes tagged Male) and $Cal_F(D)$ (corresponding of terms that are into classes tagged Female). If $Cal_M(D)$ is higher than $Cal_F(D)$ then the document is considered as written by a man otherwise it is considered as written by a woman.

IV. EVALUATION AND THE USED COLLECTION

A. Collection

The used collection is a British National Corpus (BNC), which contains 681,288 blogs, is (140 million words). The interest to use this collection is that it is tagged as genders, which allow us to have a database of comparison.

B. Evaluation

We first, implemented some quoted works (section 2), by using the same collection BNC. We calculate a measure of precision (the obtained results compared to the tagged collection). Secondly we implement our approach. The table III shows the obtained results.

Approaches	Precision
Shlomo Argamon et al. 2003 [1]	66%
Scott Nowson et al [2]	69%
Shlomo Argamon et al 2006 [3]	65%
Sumit Goswami et al [5]	67%
Arjun Mukherjee et Al [6]	68%
Sholomo Argamon et al 2007 [4]	71%
Our proposed approach	82%

Table III
THE RESULTS OF IMPLEMENTATIONS

We notice an improvement of the gender detection for more than 15 % compared to the best result obtained by Sholomo Argamon and al 2007 [4].

V. CONCLUSION

In this article, we implemented the gender detection of a blogger. We propose new characteristics based on various dictionaries that classify a blog as document written by a man or by a woman. We validate our approach using a corpus BNC that is already tagged by gender. This corpus allows us to compare the performance and pertinence or our approach. Our obtained results are relevant, where they give better results in gender detection than some related work on this topic.

As future work, it would be interesting to experiment our approach in other data collection, such as TREC Blogs track 2006. This implementation will demonstrate the robustness of our approach. We can improve our obtained score by using not just word extraction, but also sentences extraction. Finally, if we introduce a learning algorithm, we could compare our approach with learning algorithms and see if any improvement can occur.

REFERENCES

[1] Argamon Shlomo, Koppel Moshe, Fine Jonathan and Shimoni Anat R: Gender, Genre, and Writing Style in Formal Written Texts. In Journal Text, 2003.

[2] Nowson Scott, Jon, Alastair: Weblogs, genres and individual differences. In proceedings of the 27th Annual Conference of the Cognitive Science Society, 2005.

[3] Schler J., Koppel M. , Argamon S., Pennebaker J. : Effects of Age and Gender on Blogging. In proceedings of AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs, 2006.

[4] Argamon Shlomo, Koppel Moshe Pennebaker, James W., Schler Jonathan : Mining the Blogosphere: Age, gender and the varieties of self-expression. First Monday, 2007.

[5] Goswami Sumit, Sarkar Sudeshna, Rustagi Mayur: Stylometric Analysis of Bloggers Age and Gender. In proceedings of AAAI Press ICWSM, 2009.

[6] Arjun Mukherjee, Bing Liu : Improving Gender Classification of Blog Authors. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL. 2010.

[7] Cathy, Zhang, Pengyu Zhang : Predicting gender from blog posts. In proceeding of ACM, 2010.

[8] O. de Vel, M. Corney, A. Anderson, G. Mohay : Language and gender author cohort analysis of e-mail for computer forensics. In proceedings of the Second Digital Forensic Research Workshop, 2007.

[9] N. Glance, M. Hurst, and T. Tomokiyo. BlogPulse: Automated trend discovery for weblogs. In proceedings of the First Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics 2007.

[10] D. Gruhl, R. Guha, D. Liben Nowell, and A. Tomkins : Information diffusion through blogspace. In proceedings of the 13th international Conference on World Wide Web (New York, N.Y.) 2007.

[11] D.A. Huffaker and S.L. Calvert: Gender, identity, and language use in teenage blogs. In journal of Compute Mediated Communication 2007.

[12] J.W. Pennebaker, M.R. Mehl, and K. Niederhoffer: Psychological aspects of natural language use: Our words, ourselves, Annual Review of Psychology, 2006.

[13] Y. Wu and B.L. Tseng: Important Weblog identification and hot story summarization. In Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium. Menlo Park, Calif.: AAAI Press,2006.